

Robust Asymmetric Clustering

Katherine Morris* Paul D. McNicholas*[†]
Antonio Punzo[‡] Ryan P. Browne*

Abstract

Contaminated mixture models are developed for model-based clustering of data with asymmetric clusters as well as spurious points, outliers, and/or noise. Specifically, we introduce a contaminated mixture of contaminated shifted asymmetric Laplace distributions and a contaminated mixture of contaminated skew-normal distributions. In each case, mixture components have a parameter controlling the proportion of bad points (i.e., spurious points, outliers, and/or noise) and one specifying the degree of contamination. A very important feature of our approaches is that these parameters do not have to be specified *a priori*. Expectation-conditional maximization algorithms are outlined for parameter estimation and the number of components is selected using the Bayesian information criterion. The performance of our approaches is illustrated on artificial and real data.

Keywords: Contaminated mixtures, ECM algorithm, outlier detection, robust clustering, robust estimates, shifted asymmetric Laplace distribution, skew-normal distribution.

1 Introduction

Approaches have been established for clustering data with symmetric clusters when there are outlying, noisy, or spurious points, e.g., trimmed clustering (García-Escudero et al., 2008) and contaminated mixtures (Punzo and McNicholas, 2013). However, no such approach exists in situations where the clusters are asymmetric. Both the aforementioned robust clustering techniques are based on finite mixture models, i.e., model-based clustering techniques, which have burgeoned into an important subfield of cluster analysis since their use by Wolfe (1963). In this paper, we introduce methodology for robust clustering of data with asymmetric clusters. To help avoid overly verbose writing, we will use the term “bad” points to refer to outliers, spurious observations, and/or noise.

*Department of Mathematics & Statistics, University of Guelph, Guelph, Ontario, Canada, N1G 2W1.

[†]Corresponding author. E-mail: pmcnicho@uoguelph.ca

[‡]Department of Economics and Business, University of Catania, Corso Italia 55, 95129 Catania, Italy.

As discussed in Punzo and McNicholas (2013), the contaminated approach is generally preferable to the trimmed approach for clustering when bad points are present; therefore, we will focus on contaminated clustering approaches herein. The density of a p -dimensional random variable \mathbf{X} from a mixture of contaminated Gaussian distributions is given by

$$f(\mathbf{x}; \boldsymbol{\vartheta}) = \sum_{g=1}^G \pi_g \left[\lambda_g \phi(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) + (1 - \lambda_g) \phi(\mathbf{x}; \boldsymbol{\mu}_g, \eta_g \boldsymbol{\Sigma}_g) \right], \quad (1)$$

where $\pi_g > 0$ is the mixing proportion for the g th component, with $\sum_{g=1}^G \pi_g = 1$, $\lambda_g \in (0, 1)$, $\eta_g > 1$, $\phi(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the density of a p -dimensional Gaussian random variable with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$, and $\boldsymbol{\vartheta}$ denotes all model parameters. Note that α_g is restricted to an open interval to ensure identifiability (cf. Punzo and McNicholas, 2013). Note that (1) is one of many mixture of mixture models within the literature, and previous work on mixtures of mixtures can be found, for example, in Orbanz and Buhmann (2005), Di Zio et al. (2007), and Browne et al. (2012).

Over the past five years or so, there has been a substantial amount of work on clustering using finite mixtures of skewed distributions. This work can be partitioned into mixtures with component densities that have a concentration parameter and those that do not. Examples of the former include mixtures of skew- t distributions (e.g., Lin, 2010; Murray et al., 2013; Lee and McLachlan, 2014), mixtures of multivariate normal-inverse Gaussian distributions (e.g., Karlis and Santourian, 2009; Subedi and McNicholas, 2014), mixtures of variance-Gamma distributions (McNicholas et al., 2013), and mixtures of generalized hyperbolic distributions (Browne and McNicholas, 2013). There is comparatively little work, however, on clustering using skewed mixtures without a concentration parameter. Such work focuses on mixtures of skew-normal distributions (e.g., Lin, 2009) and mixture of shifted asymmetric Laplace (SAL) distributions (Franczak et al., 2012, 2013). As the name suggests, a concentration parameter, such as the degrees of freedom in the case of a skew- t distribution, controls the extent to which the density is concentrated; accordingly, a concentration parameter affects the heaviness of the tails. With this in mind, we pursue contaminated mixtures of skewed distributions that do not have a concentration parameter; specifically, we focus on mixtures of contaminated SAL distributions and mixtures of contaminated skew-normal distributions herein.

2 Methodology

2.1 Mixture of Contaminated SAL Distributions

The density of a p -variate random vector \mathbf{X} from a SAL distribution is

$$\xi(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}) = \frac{2 \exp \{(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}\}}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \left(\frac{\delta(\mathbf{x}, \boldsymbol{\mu}; \boldsymbol{\Sigma})}{2 + \boldsymbol{\alpha}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}} \right)^{\nu/2} K_{\nu}(u), \quad (2)$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is a location parameter, $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite scale matrix, $\boldsymbol{\alpha} \in \mathbb{R}^p$ is the skewness parameter, $\delta(\mathbf{x}, \boldsymbol{\mu}; \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ is the squared Mahalanobis distance between \mathbf{x} and $\boldsymbol{\mu}$, $u = \sqrt{(2 + \boldsymbol{\alpha}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}) \delta(\mathbf{x}, \boldsymbol{\mu}; \boldsymbol{\Sigma})}$, and K_ν is the modified Bessel function of the third kind with index $\nu = (2 - p)/2$. Note that the covariance of \mathbf{X} is $\text{Cov}(\mathbf{X}) = \boldsymbol{\Sigma} + \boldsymbol{\alpha} \boldsymbol{\alpha}'$, and it is important to contaminate the covariance and not just the scale matrix $\boldsymbol{\Sigma}$. To see why this is so, consider Figure 1, from which it is clear that we need to contaminate, or inflate (we only consider $\eta_g > 1$), $\text{Cov}(\mathbf{X})$.

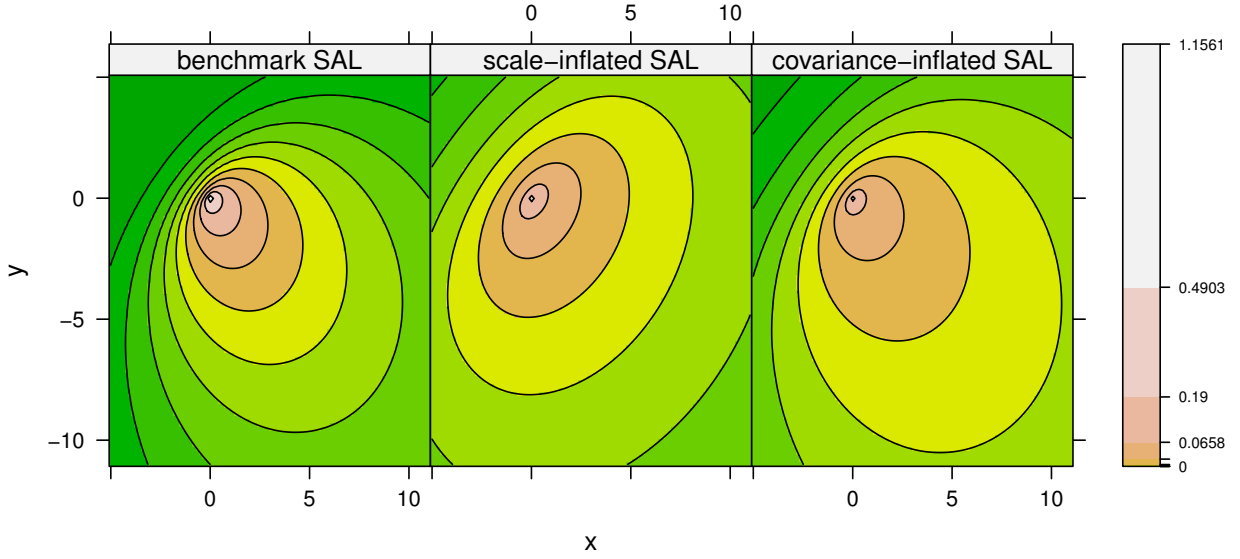


Figure 1: Contour levels for a SAL distribution (left), the same SAL distribution with the scale matrix inflated (centre), and with the covariance matrix inflated (right).

The density of a mixture of contaminated SAL distributions can be written

$$f_{\text{SAL}}(\mathbf{x}; \boldsymbol{\vartheta}) = \sum_{g=1}^G \pi_g \left[\lambda_g \xi(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\alpha}_g) + (1 - \lambda_g) \xi(\mathbf{x}; \boldsymbol{\mu}_g, \eta_g \boldsymbol{\Sigma}_g, \sqrt{\eta_g} \boldsymbol{\alpha}_g) \right], \quad (3)$$

where $\lambda_g \in (0, 1)$ and $\eta_g > 1$. As in the case of mixtures of contaminated Gaussian distributions, we restrict $\eta_g > 1$ so that the points classified into the contaminated densities $\xi(\mathbf{x}; \boldsymbol{\mu}_g, \eta_g \boldsymbol{\Sigma}_g, \sqrt{\eta_g} \boldsymbol{\alpha}_g)$ will be the bad points. It follows that the proportion of bad points in component g is $(1 - \lambda_g)$.

Suppose we observe a random sample $\mathbf{x}_1, \dots, \mathbf{x}_n$ from a mixture of contaminated SAL distributions. We introduce z_{ig} to denote component membership, so that $z_{ig} = 1$ if observation \mathbf{x}_i belongs to component g , and $z_{ig} = 0$, otherwise. Similarly, we introduce v_{ig} to indicate whether an observation is bad, so that $v_{ig} = 1$ if observation \mathbf{x}_i in component g

is not bad, and $v_{ig} = 0$, if observation \mathbf{x}_i in component g is bad. Parameter estimation for the mixture of contaminated SAL distributions is carried out using the expectation-conditional maximization (ECM) algorithm (Meng and Rubin, 1993), which is a variant of the expectation-maximization (EM) algorithm (Dempster et al., 1977). The ECM algorithm is an iterative procedure for finding maximum likelihood estimates when data are incomplete or are treated as being incomplete. Two steps are iterated until convergence is reached. In the expectation step (E-step), the expected value of the complete-data log-likelihood is computed. Then in the conditional-maximization step (CM-step), the expected value of the complete-data log-likelihood is maximized with respect to the model parameters. Extensive details on the EM algorithm and its variants are given by McLachlan and Krishnan (2008). The ECM algorithm is based on the complete-data log-likelihood, i.e., the likelihood of the observed plus the missing data.

The complete-data likelihood for the mixture of contaminated SAL distributions is

$$\mathcal{L}_{\text{SAL}}(\boldsymbol{\vartheta}) = \prod_{i=1}^n \prod_{g=1}^G \left\{ \pi_g [\lambda_g \xi(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\alpha}_g, \boldsymbol{\Sigma}_g)]^{v_{ig}} [(1 - \lambda_g) \xi(\mathbf{x}_i; \boldsymbol{\mu}_g, \sqrt{\eta_g} \boldsymbol{\alpha}_g, \eta_g \boldsymbol{\Sigma}_g)]^{(1-v_{ig})} \right\}^{z_{ig}}.$$

In their parameter estimation, Franczak et al. (2012) use the fact that a SAL random variable \mathbf{X} can be generated through the relationship

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\alpha} + \sqrt{W}\mathbf{Y}, \quad (4)$$

where $W \sim \text{Exp}(1)$ and $\mathbf{Y} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. It follows that, $\mathbf{X} \mid w \sim N(\boldsymbol{\mu} + w\boldsymbol{\alpha}, w\boldsymbol{\Sigma})$ and $W \mid \mathbf{x} \sim \text{GIG}(2 + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}, \delta(\mathbf{x}, \boldsymbol{\mu}; \boldsymbol{\Sigma}))$, cf. Franczak et al. (2012). Note that the generalized inverse Gaussian (GIG) distribution has been extensively studied within the literature, including work by Barndorff-Nielsen and Halgreen (1977), Blæsild (1978), Halgreen (1979), and Jørgensen (1982). Using (4), the complete-data log-likelihood can be written

$$\begin{aligned} l_{\text{SAL}}(\boldsymbol{\vartheta}) &= \sum_{i=1}^n \sum_{g=1}^G z_{ig} v_{ig} [\log \pi_g + \log \lambda_g + \log \phi(\mathbf{x}_i; \boldsymbol{\mu}_g + w_{ig}\boldsymbol{\alpha}_g, w_{ig}\boldsymbol{\Sigma}_g) + \log h(w_{ig})] \\ &+ \sum_{i=1}^n \sum_{g=1}^G z_{ig} (1 - v_{ig}) [\log \pi_g + \log(1 - \lambda_g) + \log \phi(\mathbf{x}_i; \boldsymbol{\mu}_g + w_{ig}\sqrt{\eta_g}\boldsymbol{\alpha}_g, w_{ig}\eta_g\boldsymbol{\Sigma}_g) + \log h(w_{ig})]. \end{aligned} \quad (5)$$

Extensive details on the ECM algorithm for the mixture of contaminated SAL distributions is given in Appendix A.

2.2 Mixture of Contaminated Skew-Normal Distributions

We employ a notation akin to that used in Lin (2009), where a flexible skew-normal mixture modelling framework was derived. A random vector \mathbf{X} that follows such a p -dimensional

skew-normal distribution with a $p \times 1$ location vector $\boldsymbol{\mu}$, a $p \times p$ positive definite scale matrix $\boldsymbol{\Sigma}$, and a $p \times p$ skewness matrix \mathbf{A} has the density function

$$\zeta(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{A}) = 2^p \phi(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Omega}) \Phi(\mathbf{A}' \boldsymbol{\Omega}^{-1}(\mathbf{x} - \boldsymbol{\mu}); \boldsymbol{\Delta}), \quad (6)$$

with $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \mathbf{A} \mathbf{A}'$ and $\boldsymbol{\Delta} = (\mathbf{I}_p + \mathbf{A}' \boldsymbol{\Sigma}^{-1} \mathbf{A})^{-1} = \mathbf{I}_p - \mathbf{A}' \boldsymbol{\Omega}^{-1} \mathbf{A}$, where \mathbf{I}_p indicates a $p \times p$ identity matrix. We will assume that \mathbf{A} is a diagonal matrix, i.e., $\mathbf{A} = \text{diag}(\boldsymbol{\alpha})$, so that the covariance of \mathbf{X} is not affected by the skewness. Also, $\Phi(\cdot)$ gives the cumulative density function of $N(\mathbf{0}, \boldsymbol{\Sigma})$.

Arelanno-Valle et al. (2007) showed that the random variable \mathbf{X} from (6) can be generated through the relationship

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{A} \boldsymbol{\tau} + \mathbf{U}, \quad (7)$$

where $\boldsymbol{\tau}$ is independently distributed as a standard half-normal distribution $\text{HN}(\mathbf{0}, \mathbf{I}_p)$, and \mathbf{U} is independently distributed as a standard normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$. It follows that $\mathbf{X} \mid \boldsymbol{\tau} \sim N(\boldsymbol{\mu} + \mathbf{A} \boldsymbol{\tau}, \boldsymbol{\Sigma})$, and the density in (6) can be written

$$\zeta(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{A}) = |2\pi \boldsymbol{\Sigma}|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{x} - (\boldsymbol{\mu} + \mathbf{A} \boldsymbol{\tau}))' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} + \mathbf{A} \boldsymbol{\tau}) \right\}. \quad (8)$$

Furthermore, Lin (2009) showed that $\boldsymbol{\tau} \mid \mathbf{X}$ can be represented via a truncated normal (TN) distribution

$$\boldsymbol{\tau} \mid \mathbf{X} \sim \text{TN}(\mathbf{A}' \boldsymbol{\Omega}^{-1}(\mathbf{x} - \boldsymbol{\mu}), \boldsymbol{\Delta}; \mathbb{R}_+^p), \quad (9)$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \mathbf{A} \mathbf{A}'$, $\boldsymbol{\Delta} = (\mathbf{I}_p + \mathbf{A}' \boldsymbol{\Sigma}^{-1} \mathbf{A})^{-1}$, and

$$\mathbb{R}_+^p = \{\mathbf{y} = (y_1, \dots, y_p)' \in \mathbb{R}^p \mid y_j > 0, j = 1, \dots, p\}.$$

The TN distribution is used here because of its computational advantages, and further details on this framework and associated properties can be found, for example, in Tallis (1961) and Lin (2009). Similar to the case of contaminated SAL mixtures (Section 2.1) we follow Punzo and McNicholas (2013) and consider the contamination scheme whereby $\text{Cov}(\mathbf{X})$ is inflated by η_g . Again, this amounts to inflating $\boldsymbol{\Sigma}_g$ to $\eta_g \boldsymbol{\Sigma}_g$ and $\boldsymbol{\alpha}_g$ to $\sqrt{\eta_g} \boldsymbol{\alpha}_g$. This leads to the following mixture of contaminated skew-normal distributions

$$f_{\text{SN}}(\mathbf{x}; \boldsymbol{\vartheta}) = \sum_{g=1}^G \pi_g [\lambda_g \zeta(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\alpha}_g, \boldsymbol{\Sigma}_g) + (1 - \lambda_g) \zeta(\mathbf{x}; \boldsymbol{\mu}_g, \sqrt{\eta_g} \boldsymbol{\alpha}_g, \eta_g \boldsymbol{\Sigma}_g)]. \quad (10)$$

where $\lambda_g \in (0, 1)$ denotes the proportion of bad points and $\eta_g > 1$. An illustration of one component from model (10), i.e., $\lambda_g \zeta(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\alpha}_g, \boldsymbol{\Sigma}_g) + (1 - \lambda_g) \zeta(\mathbf{x}; \boldsymbol{\mu}_g, \sqrt{\eta_g} \boldsymbol{\alpha}_g, \eta_g \boldsymbol{\Sigma}_g)$, is given in Figure 2.

Like the mixture of contaminated SAL distributions, an ECM algorithm is used for parameter estimation for the mixture of contaminated skew-normal distributions. Ignoring

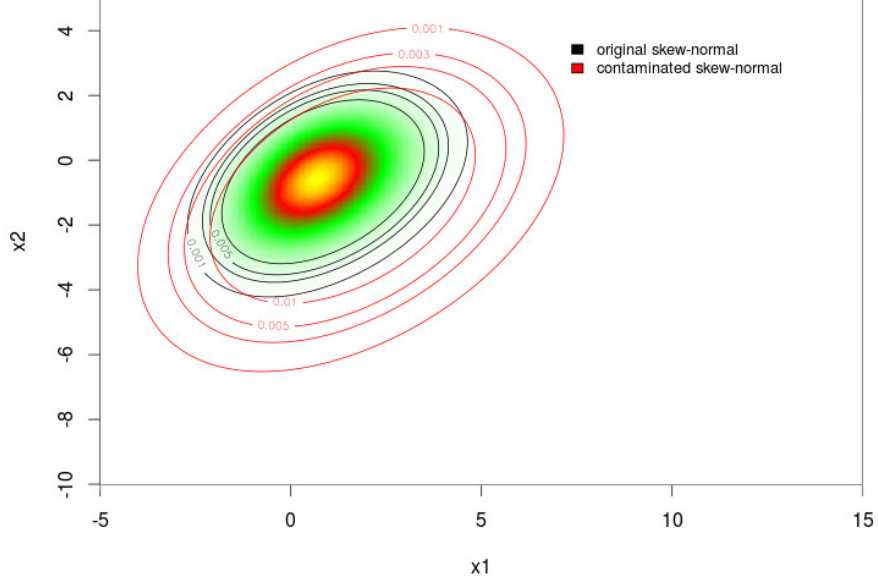


Figure 2: Contour plot of a contaminated skew-normal mixture component.

additive constants, the complete-data log-likelihood can be written

$$\begin{aligned}
l_{\text{SN}}(\boldsymbol{\vartheta}) = & \sum_{i=1}^n \sum_{g=1}^G z_{ig} v_{ig} \left\{ \log \pi_g + \log \lambda_g - \frac{1}{2} \log |\boldsymbol{\Sigma}_g^{-1}| \right. \\
& \left. - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_g - \mathbf{A}_g \boldsymbol{\tau}_{ig})' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g - \mathbf{A}_g \boldsymbol{\tau}_{ig}) - \frac{1}{2} \boldsymbol{\tau}_{ig}' \boldsymbol{\tau}_{ig} \right\} \\
& + \sum_{i=1}^n \sum_{g=1}^G z_{ig} (1 - v_{ig}) \left\{ \log \pi_g + \log(1 - \lambda_g) - \frac{1}{2} \log |(\eta_g \boldsymbol{\Sigma}_g)^{-1}| \right. \\
& \left. - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_g - \sqrt{\eta_g} \mathbf{A}_g \tilde{\boldsymbol{\tau}}_{ig})' (\eta_g \boldsymbol{\Sigma}_g)^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g - \sqrt{\eta_g} \mathbf{A}_g \tilde{\boldsymbol{\tau}}_{ig}) - \frac{1}{2} \tilde{\boldsymbol{\tau}}_{ig}' \tilde{\boldsymbol{\tau}}_{ig} \right\},
\end{aligned}$$

where $\tilde{\boldsymbol{\tau}}$ is the analogue of $\boldsymbol{\tau}$ when $\boldsymbol{\Sigma}_g$ and \mathbf{A}_g are replaced by $\eta_g \boldsymbol{\Sigma}_g$ and $\sqrt{\eta_g} \mathbf{A}_g$, respectively. Further details, including parameter updates, for this ECM algorithm are given in Appendix B.

3 Applications

3.1 Model Selection and Performance Assessment

For each example, we fit our contaminated mixtures for values of G between G_{true} and $G_{\text{true}} + 2$, and use the Bayesian information criterion (BIC; Schwarz, 1978) to select the best model (i.e., to select the value of G). Note that the BIC has frequently been used throughout the literature for mixture model selection (e.g., Dasgupta and Raftery, 1998; Fraley and Raftery, 2002; McNicholas and Murphy, 2008) and is given by

$$\text{BIC} = 2l(\mathbf{x}, \hat{\boldsymbol{\theta}}) - \rho \log n,$$

where ρ is the number of free parameters and $l(\mathbf{x}, \hat{\boldsymbol{\theta}})$ is the maximized (observed) log-likelihood.

Predicted classifications can be compared to true class labels using the adjusted Rand index (ARI; Hubert and Arabie, 1985), which is the Rand index (Rand, 1971) corrected for chance agreement. The Rand index is based on pairwise agreements and takes a value between 0 and 1, where 1 indicates perfect agreement between two partitions. The correction that leads to the ARI accounts for the fact that random classification is expected to result in some correct agreements; accordingly, the ARI has an expected value of 0 under random classification and, as with the Rand index, perfect classification corresponds to a value of 1. Negative ARI values are possible and indicate classification results worse than would be expected by random classification.

3.2 Simulated data

We generated $n = 180$ data points from a SAL mixture with two components of equal size ($n_1 = n_2 = 90$) using the relationship in (4). Similarly, we simulated $n = 180$ data points from a skew-normal mixture with two components of equal size ($n_1 = n_2 = 90$) using the relationship in (7). The specific parameters used to simulate these data sets appear in Table 1. We then artificially added 20 bad points in each case.

Table 1: Parameters used to simulate data for the contaminated SAL and skew-normal mixtures.

Component 1	Component 2	Note
$W_1 \sim \text{Exp}(1)$	$W_2 \sim \text{Exp}(1)$	SAL mixtures
$\boldsymbol{\tau}_1 \sim \text{HN}(0, 1)$	$\boldsymbol{\tau}_2 \sim \text{HN}(0, 1)$	Skew-normal mixtures
$\boldsymbol{\alpha}_1 = (1, 1)'$	$\boldsymbol{\alpha}_2 = (-1, -1)'$	
$\boldsymbol{\mu}_1 = (1, -2)'$	$\boldsymbol{\mu}_2 = (5, -5)'$	
$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	

We generated a total of 100 such data sets in each case (i.e., SAL mixtures and skew-normal mixtures). We fitted our contaminated SAL mixtures to each data set that arose from a SAL mixture, and we fitted our contaminated skew-normal mixtures to each data set that arose from a skew-normal mixture. One example of a fitted model, in each case, is given in Figure 3. Overall, classification performance was very good in each case. Specifically, in the SAL case the average ARI was 0.8199 with a standard deviation of 0.0495, and for the skew-normal case, the average ARI was 0.8955 with a standard deviation of 0.0349. Many of the misclassifications arose when bad points were close to components and so classified as belonging to components (e.g., Figure 3).

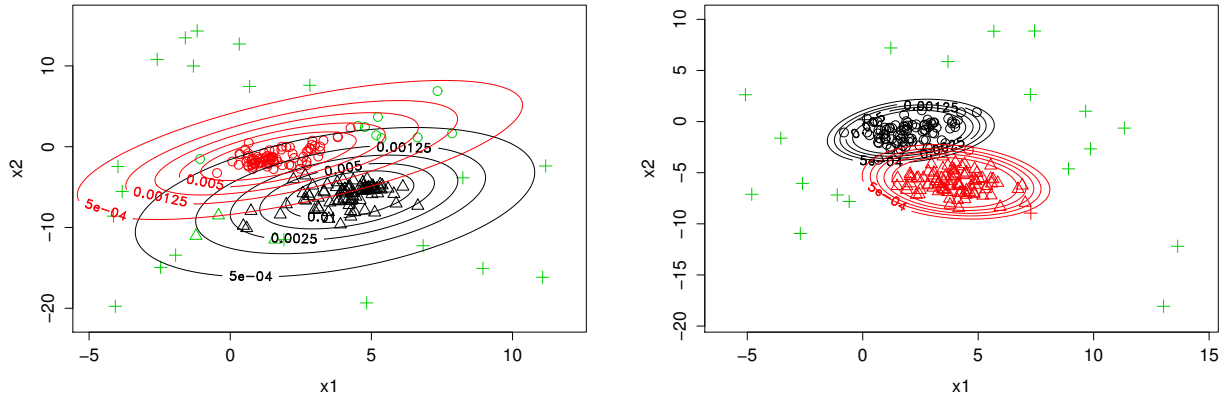


Figure 3: Scatterplots for a fitted contaminated SAL mixture (left) and a fitted contaminated skew-normal mixture (right) applied to simulated data, where colour indicates the estimated component membership and symbol (i.e., \circ , \triangle , or $+$) indicates true membership, with the symbol $+$ representing bad points.

3.3 Real data

In this section we consider four real data sets.

- Cook and Weisberg (1994) provided eleven body and blood measurements collected from athletes at the Australian Institute of Sport (AIS). We used the data available from the R package `sn` (Azzalini, 2013). There are 102 male and 100 female athletes.
- Forina et al. (1986) recorded chemical and physical properties for three types of Italian wines: Barolo, Grignolino, and Barbera. The data available from the R package `gclus` (Hurley, 2010) contain thirteen variables and 178 samples.
- Flury and Riedwyl (1988) present six measurements taken from Swiss banknotes, which are available through the R package `gclus`. There are genuine and counterfeit notes.

- Flury (1997, Table 5.3.7) discuss seven measurements of female voles from two species (*Microtus californicus* and *Microtus ochrogaster*, originally studied by Airolidi and Hoffmann, 1984).

The bank notes data have two natural outliers, and for the other three data sets we introduce an artificial outlier. To this end, we pick an observation and add to it suitable multiples of the first and/or second principal component vectors (Figure 4). The PCs are determined using robust principal component analysis (ROBPCA; Hubert et al., 2005), as implemented in the R package `rrcov` (Todorov and Filzmoser, 2009).

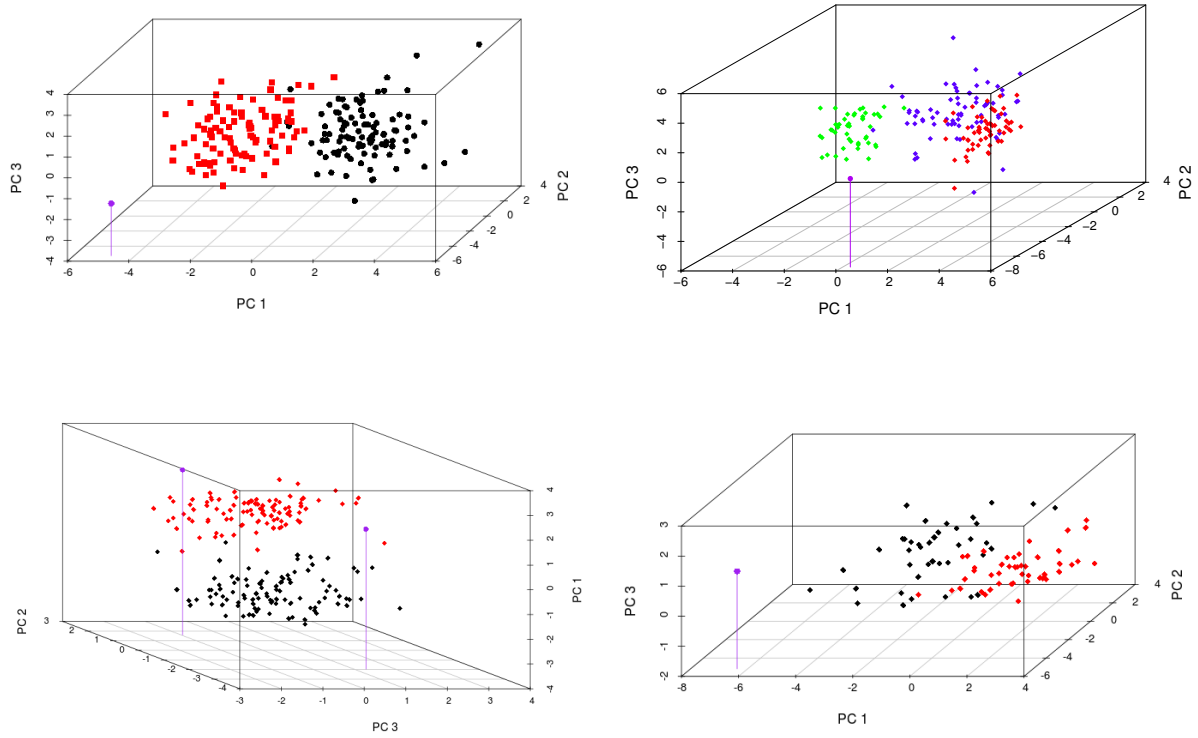


Figure 4: Principal components for the real data sets — AIS (top, left), wine (top, right), bank notes (bottom, left), and voles (bottom, right) — considered with true classes indicated and outliers clearly highlighted.

For each data set, we applied mixtures of contaminated SAL distributions as well as mixtures of contaminated skew-normal distributions for values of G between G_{true} . Both approaches performed well on all real data sets, giving high ARI values in each case (Table 2).

Table 2: Summary of clustering results for the best model, in terms of BIC, fitted to the real data using contaminated mixtures.

Data	Model	ARI
AIS	Contaminated SAL mixtures	0.8932
	Contaminated skew-normal mixtures	0.8748
Bank	Contaminated SAL mixtures	0.9205
	Contaminated skew-normal mixtures	0.9602
Voles	Contaminated SAL mixtures	0.9540
	Contaminated skew-normal mixtures	0.9540
Wine	Contaminated SAL mixtures	0.9226
	Contaminated skew-normal mixtures	0.9172

4 Conclusion

Two approaches have been presented for the previously unsolved problem of clustering data with bad points and asymmetric clusters. Both approaches are based on contaminated mixture models; specifically, a mixture of contaminated SAL distributions and a mixture of contaminated skew-normal distributions are developed. As such, this work is an extension of the contaminated Gaussian mixtures approach of Punzo and McNicholas (2013) to SAL and skew-normal mixtures, respectively. In both cases, an ECM algorithm is outlined for model selection and details are given in appendices. One major advantage of our approach — and the contaminated approach in general (cf. Punzo and McNicholas, 2013) — is that we do not need to specify the proportion of outliers *a priori*. Both methods were applied to simulated and real data where they gave very good classification performance. Future work will focus on contamination of skewed mixtures where a concentration parameter is present in each component, as well as on introducing parsimony into the models introduced herein.

Acknowledgement

This work was supported by an Ontario Graduate Scholarship (Morris), an Early Researcher Award from the Government of Ontario (McNicholas), and a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC; McNicholas). The computing equipment used was provided through a Research Tools and Instruments Grant from NSERC. The authors are grateful to Prof. Tsung-I Lin for providing code to help implement the skew-normal models used herein.

References

Airoidi, J.-P. and R. S. Hoffmann (1984). Age variation in voles (*Microtus californicus*, *M. ochrogaster*) and its significance for systematic studies. *Occasional papers of the Museum*

- of *Natural History*, University of Kansas, Lawrence KS 111, 1–45.
- Arelanno-Valle, R. B., H. Bolfarine, and V. Lachos (2007). Bayesian inference for skew-normal linear mixed models. *Journal of Applied Statistics* 34, 663–682.
- Azzalini, A. (2013). *R package **sn**: The skew-normal and skew-t distributions (version 0.4-18)*. Università di Padova, Italia.
- Barndorff-Nielsen, O. and C. Halgreen (1977). Infinite divisibility of the hyperbolic and generalized inverse Gaussian distributions. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* 38, 309–311.
- Blæsild, P. (1978). The shape of the generalized inverse Gaussian and hyperbolic distributions. Research Report 37, Aarhus University, Denmark, Department of Theoretical Statistics.
- Browne, R. P. and P. D. McNicholas (2013). A mixture of generalized hyperbolic distributions. arXiv:1305.1036.
- Browne, R. P., P. D. McNicholas, and M. D. Sparling (2012). Model-based learning using a mixture of mixtures of Gaussian and uniform distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(4), 814–817.
- Cook, R. D. and S. Weisberg (1994). *An Introduction to Regression Graphics*. New York: John Wiley and Sons.
- Dasgupta, A. and A. Raftery (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association* 93(441), 294–302.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* 39(1), 1–38.
- Di Zio, M., U. Guarnera, and R. Rocci (2007). A mixture of mixture models for a classification problem: The unity measure error. *Computational Statistics & Data analysis* 51(5), 2573–2585.
- Flury, B. and H. Riedwyl (1988). *Multivariate Statistics: A Practical Approach*. Cambridge University Press.
- Flury, B. D. (1997). *A First Course in Multivariate Statistics*. New York: Springer.
- Forina, M., C. Armanino, M. Castino, and M. Ubigli (1986). Multivariate data analysis as a discriminating method of the origin of wines. *Vitis* 25, 189–201.

- Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97(458), 611–631.
- Franczak, B., R. P. Browne, and P. D. McNicholas (2012). Mixtures of shifted asymmetric Laplace distributions. arXiv:1207.1727v3.
- Franczak, B. C., P. D. McNicholas, R. P. Browne, and P. M. Murray (2013). Parsimonious shifted asymmetric Laplace mixtures. arXiv:1311.0317.
- García-Escudero, L. A., A. Gordaliza, C. Matrán, and A. Mayo-Iscar (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics* 36(3), 1324–1345.
- Halgreen, C. (1979). Self-decomposability of the generalized inverse Gaussian and hyperbolic distributions. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* 47, 13–18.
- Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification* 2, 193–218.
- Hubert, M., P. J. Rousseeuw, and K. Vanden Branden (2005). ROBPCA: a new approach to robust principal components analysis. *Technometrics* 47, 64–79.
- Hurley, C. (2010). *gclus: Clustering Graphics*. R package version 1.3.
- Jørgensen, B. (1982). *Statistical Properties of the Generalized Inverse Gaussian Distribution*. New York: Springer-Verlag.
- Karlis, D. and A. Santourian (2009). Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing* 19(1), 73–83.
- Lee, S. and G. J. McLachlan (2014). Finite mixtures of multivariate skew t -distributions: some recent and new results. *Statistics and Computing* 24(2), 181–202.
- Lin, T.-I. (2009). Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis* 100, 257–265.
- Lin, T.-I. (2010). Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing* 20(3), 343–356.
- McLachlan, G. J. and T. Krishnan (2008). *The EM Algorithm and Extensions* (2nd ed.). New York: Wiley.
- McNicholas, P. D. and T. B. Murphy (2008). Parsimonious Gaussian mixture models. *Statistics and Computing* 18, 285–296.
- McNicholas, S. M., P. D. McNicholas, and R. P. Browne (2013). Mixtures of variance-gamma distributions. Arxiv preprint arXiv:1309.2695.

- Meng, X. L. and D. B. Rubin (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 80, 267–278.
- Murray, P. M., R. P. Browne, and P. D. McNicholas (2013). Mixtures of skew- t factor analyzers. arXiv:1305.4301v2.
- Orbanz, P. and J. M. Buhmann (2005). SAR images as mixtures of Gaussian mixtures. *IEEE International Conference on Image Processing 2*, 209–212.
- Punzo, A. and P. D. McNicholas (2013). Outlier detection via parsimonious mixtures of contaminated gaussian distributions. arXiv: 1305.4669v1.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66, 846–850.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.
- Subedi, S. and P. D. McNicholas (2014). Variational Bayes approximations for clustering via mixtures of normal inverse Gaussian distributions. *Advances in Data Analysis and Classification*. To appear.
- Tallis, G. (1961). The moment generating function of the truncated multi-normal distribution. *Journal of the Royal Statistical Society: Series B* 23, 223–229.
- Todorov, V. and P. Filzmoser (2009). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software* 32(3), 1–47.
- Wolfe, J. H. (1963). Object cluster analysis of social areas. Master’s thesis, University of California, Berkeley.

A EM Algorithm for Mixture of Contaminated SAL Distributions

One can simplify (5) to

$$\begin{aligned}
l_{\text{SAL}}(\boldsymbol{\vartheta}) &= \sum_{i=1}^n \sum_{g=1}^G z_{ig} v_{ig} \log \pi_g + \sum_{i=1}^n \sum_{g=1}^G z_{ig} v_{ig} \log \lambda_g \\
&+ \sum_{i=1}^n \sum_{g=1}^G z_{ig} v_{ig} \log \left(\frac{1}{(2\pi)^{p/2} |w_{ig} \boldsymbol{\Sigma}_g|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_g - w_{ig} \boldsymbol{\alpha}_g)' \right. \right. \\
&\times \left. \left. (w_{ig} \boldsymbol{\Sigma}_g)^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g - w_{ig} \boldsymbol{\alpha}_g) \exp(-w_{ig}) \right\} \right) \\
&+ \sum_{i=1}^n \sum_{g=1}^G z_{ig} (1 - v_{ig}) \log(\pi_g) + \sum_{i=1}^n \sum_{g=1}^G z_{ig} (1 - v_{ig}) \log(1 - \lambda_g) \\
&+ \sum_{i=1}^n \sum_{g=1}^G z_{ig} (1 - v_{ig}) \log \left(\frac{1}{(2\pi)^{p/2} |w_{ig} \eta_g \boldsymbol{\Sigma}_g|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_g - w_{ig} \sqrt{\eta_g} \boldsymbol{\alpha}_g)' \right. \right. \\
&\times \left. \left. (w_{ig} \eta_g \boldsymbol{\Sigma}_g)^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g - w_{ig} \sqrt{\eta_g} \boldsymbol{\alpha}_g) \exp(-w_{ig}) \right\} \right).
\end{aligned} \tag{11}$$

We employ the ECM algorithm to find the maximum likelihood estimates in (11). At the E-step of the algorithm the Q-function, $Q(\boldsymbol{\vartheta}|\hat{\boldsymbol{\vartheta}}) = E[l(\boldsymbol{\vartheta})|\mathbf{x}, \hat{\boldsymbol{\vartheta}}]$ is computed. This Q-function denotes the conditional expectation of (11) given the observed data \mathbf{x} and the current estimated parameters $\hat{\boldsymbol{\vartheta}}$. For simplicity we let $l_{\text{SAL}}(\boldsymbol{\vartheta}) = l_1(\boldsymbol{\vartheta}) + l_2(\boldsymbol{\vartheta})$, where $l_1(\boldsymbol{\vartheta})$ contains the first three terms from (11) and $l_2(\boldsymbol{\vartheta})$ contains the last three terms from (11). Then we can write the Q-function as

$$Q(\boldsymbol{\vartheta}|\hat{\boldsymbol{\vartheta}}) = E[l_1(\boldsymbol{\vartheta})] + E[l_2(\boldsymbol{\vartheta})], \tag{12}$$

thus for the first term in the Q-function (12) we obtain

$$\begin{aligned}
E[l_1(\boldsymbol{\vartheta})] &= \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} \hat{v}_{ig} \log(\pi_g) + \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} \hat{v}_{ig} \log(\lambda_g) \\
&- \sum_{i=1}^n \sum_{g=1}^G \frac{\hat{z}_{ig} \hat{v}_{ig} p}{2} \log(2\pi) + \sum_{i=1}^n \sum_{g=1}^G \frac{\hat{z}_{ig} \hat{v}_{ig}}{2} \log(|\boldsymbol{\Sigma}_g^{-1}|) \\
&- \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} \hat{v}_{ig} E[W_{ig}^{-1} | \mathbf{x}_i, z_{ig} = 1] (\mathbf{x}_i - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g) \\
&- \frac{p}{2} \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} \hat{v}_{ig} E[\log(W_{ig}) | \mathbf{x}_i, z_{ig} = 1] \\
&+ \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} \hat{v}_{ig} (\mathbf{x}_i - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} \boldsymbol{\alpha}_g + \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} \hat{v}_{ig} \boldsymbol{\alpha}_g' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g) \\
&- \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} \hat{v}_{ig} E[W_{ig} | \mathbf{x}_i, z_{ig} = 1] \boldsymbol{\alpha}_g' \boldsymbol{\Sigma}_g^{-1} \boldsymbol{\alpha}_g \\
&- \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} \hat{v}_{ig} E[W_{ig} | \mathbf{x}_i, z_{ig} = 1].
\end{aligned} \tag{13}$$

Since the terms $-\frac{p}{2} \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} \hat{v}_{ig} E[\log(W_{ig}) | \mathbf{x}_i, z_{ig} = 1]$ and $-\sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} \hat{v}_{ig} \times E[W_{ig} | \mathbf{x}_i, z_{ig} = 1]$ are constant with respect to the model parameters, we can omit them from our calculations. The second term in the Q-function (12) has the following structure

$$\begin{aligned}
E[l_2(\boldsymbol{\vartheta})] &= \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig}(1 - \hat{v}_{ig}) \log(1 - \lambda_g) + \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig}(1 - \hat{v}_{ig}) \log(\pi_g) \\
&- \sum_{i=1}^n \sum_{g=1}^G \frac{\hat{z}_{ig}(1 - \hat{v}_{ig})p}{2} \log(2\pi) + \sum_{i=1}^n \sum_{g=1}^G \frac{\hat{z}_{ig}(1 - \hat{v}_{ig})}{2} \log(|(\eta_g \boldsymbol{\Sigma}_g)^{-1}|) \\
&- \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\tilde{W}_{ig}^{-1} | \mathbf{x}_i, z_{ig} = 1] (\mathbf{x}_i - \boldsymbol{\mu}_g)' (\eta_g \boldsymbol{\Sigma}_g)^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g) \\
&- \frac{p}{2} \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\log(\tilde{W}_{ig}) | \mathbf{x}_i, z_{ig} = 1] \\
&+ \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig}(1 - \hat{v}_{ig}) (\mathbf{x}_i - \boldsymbol{\mu}_g)' (\eta_g \boldsymbol{\Sigma}_g)^{-1} (\sqrt{\eta_g} \boldsymbol{\alpha}_g) \\
&+ \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig}(1 - \hat{v}_{ig}) (\sqrt{\eta_g} \boldsymbol{\alpha}_g)' (\eta_g \boldsymbol{\Sigma}_g)^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g) \\
&- \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\tilde{W}_{ig} | \mathbf{x}_i, z_{ig} = 1] (\sqrt{\eta_g} \boldsymbol{\alpha}_g)' (\eta_g \boldsymbol{\Sigma}_g)^{-1} (\sqrt{\eta_g} \boldsymbol{\alpha}_g) \\
&- \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\tilde{W}_{ig} | \mathbf{x}_i, z_{ig} = 1]. \tag{14}
\end{aligned}$$

Here \tilde{W} is the analogue of W when $\boldsymbol{\Sigma}_g$ becomes $\eta_g \boldsymbol{\Sigma}_g$ and $\boldsymbol{\alpha}_g$ becomes $\sqrt{\eta_g} \boldsymbol{\alpha}_g$. Once again two terms can be omitted from the calculations, namely $-\frac{p}{2} \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig}(1 - \hat{v}_{ig}) E[\log(\tilde{W}_{ig}) | \mathbf{x}_i, z_{ig} = 1]$ and $-\sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} \hat{v}_{ig} E[\tilde{W}_{ig} | \mathbf{x}_i, z_{ig} = 1]$, because they are constant with respect to the model parameters.

Now, at the E-step of the ECM algorithm we compute the update for z_{ig}

$$\hat{z}_{ig} = \frac{\pi_g [\lambda_g \xi(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\alpha}_g) + (1 - \lambda_g) \xi(\mathbf{x}; \boldsymbol{\mu}_g, \eta_g \boldsymbol{\Sigma}_g, \sqrt{\eta_g} \boldsymbol{\alpha}_g)]}{\sum_{g=1}^G \pi_g [\lambda_g \xi(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\alpha}_g) + (1 - \lambda_g) \xi(\mathbf{x}; \boldsymbol{\mu}_g, \eta_g \boldsymbol{\Sigma}_g, \sqrt{\eta_g} \boldsymbol{\alpha}_g)]} \tag{15}$$

and also derive the update for v_{ig}

$$\hat{v}_{ig} = \frac{\lambda_g \xi(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\alpha}_g)}{\lambda_g \xi(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\alpha}_g) + (1 - \lambda_g) \xi(\mathbf{x}; \boldsymbol{\mu}_g, \eta_g \boldsymbol{\Sigma}_g, \sqrt{\eta_g} \boldsymbol{\alpha}_g)}. \tag{16}$$

At the first CM-step of the ECM algorithm we update π_g and λ_g with $\hat{\pi}_g = \frac{\sum_{i=1}^n \hat{z}_{ig}}{n}$ and $\hat{\lambda}_g = \frac{\sum_{i=1}^n \hat{z}_{ig} \hat{v}_{ig}}{\sum_{i=1}^n \hat{z}_{ig}}$, respectively. Then we use the Q -function (12) to compute parameter

updates for $\boldsymbol{\mu}_g$, $\boldsymbol{\Sigma}_g$, \mathbf{A}_g and η_g . In what follows we introduce some notation to simplify the calculation of these updates. First, we set

$$\begin{aligned}
E_{1ig} &:= E[W_{ig} | \mathbf{x}_i, z_{ig} = 1] = \sqrt{\frac{a}{b}} K_\nu(u), \\
E_{2ig} &:= E[W_{ig}^{-1} | \mathbf{x}_i, z_{ig} = 1] = \sqrt{\frac{b}{a}} K_\nu(u) - \frac{2\nu}{a}, \\
E_{3ig} &:= E[\tilde{W}_{ig} | \mathbf{x}_i, z_{ig} = 1] = \sqrt{\frac{\tilde{a}}{\tilde{b}}} K_\nu(\tilde{u}), \\
E_{4ig} &:= E[\tilde{W}_{ig}^{-1} | \mathbf{x}_i, z_{ig} = 1] = \sqrt{\frac{\tilde{b}}{\tilde{a}}} K_\nu(\tilde{u}) - \frac{2\nu}{\tilde{a}},
\end{aligned} \tag{17}$$

where $a = (\mathbf{x}_i - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g)$ and $b = 2 + \boldsymbol{\alpha}_g' \boldsymbol{\Sigma}_g^{-1} \boldsymbol{\alpha}_g$, with \tilde{a} and \tilde{b} being their analogues when $\boldsymbol{\Sigma}_g^{-1}$ becomes $(\eta_g \boldsymbol{\Sigma}_g)^{-1}$ and $\boldsymbol{\alpha}_g$ becomes $\sqrt{\eta_g} \boldsymbol{\alpha}_g$. Also, we let $u = \sqrt{ab}$ and $\tilde{u} = \sqrt{\tilde{a}\tilde{b}}$. We use the following

$$\begin{aligned}
A &= \sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} E_{2ig} + \frac{1 - \hat{v}_{ig}}{\eta_g} E_{4ig}), \\
B &= \sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} E_{1ig} + (1 - \hat{v}_{ig}) E_{3ig}), \\
C &= \sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\eta_g}), \\
D &= \sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\eta_g}}),
\end{aligned} \tag{18}$$

to simplify the notation in the calculations. It turns out that the update for $\boldsymbol{\mu}_g$ becomes

$$\hat{\boldsymbol{\mu}}_g = \frac{B \left(\sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} E_{2ig} + \frac{1 - \hat{v}_{ig}}{\eta_g} E_{4ig}) \mathbf{x}_i \right) - C \left(\sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\eta_g}) \mathbf{x}_i \right)}{BA - C^2}, \tag{19}$$

and the update for $\boldsymbol{\alpha}_g$ is

$$\hat{\boldsymbol{\alpha}}_g = \frac{A \left(\sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\eta_g}}) \mathbf{x}_i \right) - D \left(\sum_{i=1}^n \hat{z}_{ig} (\hat{v}_{ig} E_{2ig} + \frac{1 - \hat{v}_{ig}}{\eta_g} E_{4ig}) \mathbf{x}_i \right)}{BA - D^2}, \tag{20}$$

while the update for Σ_g becomes

$$\begin{aligned}
\hat{\Sigma}_g &= \frac{1}{\sum_{i=1}^n \hat{z}_{ig}} A(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)' \\
&- \frac{1}{\sum_{i=1}^n \hat{z}_{ig}} \sum_{i=1}^n \hat{z}_{ig} \left(\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\eta_g}} \right) (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g) \hat{\boldsymbol{\alpha}}_g' \\
&- \frac{1}{\sum_{i=1}^n \hat{z}_{ig}} \sum_{i=1}^n \hat{z}_{ig} \left(\hat{v}_{ig} + \frac{1 - \hat{v}_{ig}}{\sqrt{\eta_g}} \right) \hat{\boldsymbol{\alpha}}_g (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)' \\
&+ \frac{1}{\sum_{i=1}^n \hat{z}_{ig}} \boldsymbol{\alpha}_g \boldsymbol{\alpha}_g' B.
\end{aligned} \tag{21}$$

At the second CM-step of the ECM algorithm we update the contamination factor η_g . Looking at the Q-function in (12) we observe that there are two possible ways of updating η_g .

1. Use the R function `optim()` to find a maximum for Q , as η_g varies in the interval $(1, \eta_g^{\max})$, where η_g^{\max} is the maximum allowed value for η_g , by considering

$$\begin{aligned}
Q &= \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} (1 - \hat{v}_{ig}) \left(-\frac{p}{2} \log(\eta_g) \right) \\
&- \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} (1 - \hat{v}_{ig}) E_{4ig}(\mathbf{x}_i - \boldsymbol{\mu}_g)' (\eta_g \Sigma_g)^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g) \\
&+ \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} (1 - \hat{v}_{ig}) (\mathbf{x}_i - \boldsymbol{\mu}_g)' (\eta_g \Sigma_g)^{-1} (\sqrt{\eta_g} \boldsymbol{\alpha}_g) \\
&+ \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} (1 - \hat{v}_{ig}) (\sqrt{\eta_g} \boldsymbol{\alpha}_g)' (\eta_g \Sigma_g)^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g) \\
&+ \text{terms independent of } \eta_g.
\end{aligned} \tag{22}$$

2. Solve the equation $\frac{\partial Q}{\partial \eta_g} = 0$ for η_g directly,

$$\begin{aligned}
0 &= \eta_g p \sum_{i=1}^n \frac{\hat{z}_{ig} (1 - \hat{v}_{ig})}{2} + \eta_g^{1/2} \left[\frac{1}{4} \sum_{i=1}^n \hat{z}_{ig} (1 - \hat{v}_{ig}) (\mathbf{x}_i - \boldsymbol{\mu}_g)' \Sigma_g^{-1} \boldsymbol{\alpha}_g \right. \\
&+ \left. \frac{1}{4} \sum_{i=1}^n \hat{z}_{ig} (1 - \hat{v}_{ig}) \boldsymbol{\alpha}_g' \Sigma_g^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g) \right] \\
&- \frac{1}{2} \sum_{i=1}^n \hat{z}_{ig} (1 - \hat{v}_{ig}) E_{4ig}(\mathbf{x}_i - \boldsymbol{\mu}_g)' \Sigma_g^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g),
\end{aligned} \tag{23}$$

which is a quadratic equation of the form $S\eta_g + T\eta_g^{1/2} + U = 0$ that can be easily solved for η_g (where S, T, U are constants).

We found it beneficial to implement both methods of computing the η_g update because they can be used interchangeably in case of numerical errors.

B EM Algorithm for Mixture of Contaminated Skew-Normal Distributions

The E-step of the algorithm relies on the computation of the Q-function, $Q(\boldsymbol{\vartheta}|\hat{\boldsymbol{\vartheta}}) = E[l(\boldsymbol{\vartheta})|\mathbf{x}, \hat{\boldsymbol{\vartheta}}]$. This Q-function denotes the conditional expectation of (10) given the observed data \mathbf{x} and the current estimated parameters $\hat{\boldsymbol{\vartheta}}$. The conditional expectations of the terms $-\frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G z_{ig} v_{ig} \boldsymbol{\tau}_{ig}' \boldsymbol{\tau}_{ig}$ and $-\frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G z_{ig} (1 - v_{ig}) \tilde{\boldsymbol{\tau}}_{ig}' \tilde{\boldsymbol{\tau}}_{ig}$ can be omitted because they do not contain any parameters. For the computations involved in the Q-function we adopt the following notation

$$\begin{aligned} E[\boldsymbol{\tau}_{ig}|\mathbf{x}_i, z_{ig} = 1] &= \boldsymbol{\eta}_{ig}, & E[\boldsymbol{\tau}_{ig} \boldsymbol{\tau}_{ig}'|\mathbf{x}_i, z_{ig} = 1] &= \boldsymbol{\psi}_{ig} \quad \text{and} \\ E[\tilde{\boldsymbol{\tau}}_{ig}|\mathbf{x}_i, z_{ig} = 1] &= \tilde{\boldsymbol{\eta}}_{ig}, & E[\tilde{\boldsymbol{\tau}}_{ig} \tilde{\boldsymbol{\tau}}_{ig}'|\mathbf{x}_i, z_{ig} = 1] &= \tilde{\boldsymbol{\psi}}_{ig}, \end{aligned} \quad (24)$$

where $\boldsymbol{\eta}_{ig}, \tilde{\boldsymbol{\eta}}_{ig}, \boldsymbol{\psi}_{ig}, \tilde{\boldsymbol{\psi}}_{ig}$ are all implicit functions of the parameters $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \mathbf{A}_g)$ and η_g where required, and can be calculated numerically using the properties of the TN distribution, as outlined in Lin (2009). The Q-function for the contaminated SN mixture model is defined as

$$\begin{aligned} Q &= \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} \hat{v}_{ig} \left\{ \log(\lambda_g) + \log(\pi_g) + \frac{1}{2} \log |\boldsymbol{\Sigma}_g^{-1}| \right. \\ &\quad - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_g - \mathbf{A}_g \boldsymbol{\eta}_{ig})' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g - \mathbf{A}_g \boldsymbol{\eta}_{ig}) \\ &\quad \left. - \frac{1}{2} \text{tr}(\boldsymbol{\Sigma}_g^{-1} \mathbf{A}_g (\boldsymbol{\psi}_{ig} - \boldsymbol{\eta}_{ig} \boldsymbol{\eta}_{ig}') \mathbf{A}_g') \right\} \\ &\quad + \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} (1 - \hat{v}_{ig}) \left\{ \log(1 - \lambda_g) + \log(\pi_g) + \frac{1}{2} \log |(\eta_g \boldsymbol{\Sigma}_g)^{-1}| \right. \\ &\quad - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_g - \sqrt{\eta_g} \mathbf{A}_g \tilde{\boldsymbol{\eta}}_{ig})' (\eta_g \boldsymbol{\Sigma}_g)^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g - \sqrt{\eta_g} \mathbf{A}_g \tilde{\boldsymbol{\eta}}_{ig}) \\ &\quad \left. - \frac{1}{2} \text{tr}((\eta_g \boldsymbol{\Sigma}_g)^{-1} \sqrt{\eta_g} \mathbf{A}_g (\tilde{\boldsymbol{\psi}}_{ig} - \tilde{\boldsymbol{\eta}}_{ig} \tilde{\boldsymbol{\eta}}_{ig}') (\sqrt{\eta_g} \mathbf{A}_g)') \right\}. \end{aligned} \quad (25)$$

At the E-step of the ECM algorithm we also update the z_{ig} with

$$\hat{z}_{ig} = \frac{\pi_g [\lambda_g \zeta(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\alpha}_g, \boldsymbol{\Sigma}_g) + (1 - \lambda_g) \zeta(\mathbf{x}; \boldsymbol{\mu}_g, \sqrt{\eta_g} \boldsymbol{\alpha}_g, \eta_g \boldsymbol{\Sigma}_g)]}{\sum_{g=1}^G \pi_g [\lambda_g \zeta(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\alpha}_g, \boldsymbol{\Sigma}_g) + (1 - \lambda_g) \zeta(\mathbf{x}; \boldsymbol{\mu}_g, \sqrt{\eta_g} \boldsymbol{\alpha}_g, \eta_g \boldsymbol{\Sigma}_g)]} \quad (26)$$

and similarly update the v_{ig} with

$$\hat{v}_{ig} = \frac{\lambda_g \zeta(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\alpha}_g, \boldsymbol{\Sigma}_g)}{\lambda_g \zeta(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\alpha}_g, \boldsymbol{\Sigma}_g) + (1 - \lambda_g) \zeta(\mathbf{x}; \boldsymbol{\mu}_g, \sqrt{\eta_g} \boldsymbol{\alpha}_g, \eta_g \boldsymbol{\Sigma}_g)}. \quad (27)$$

At the first CM-step of the ECM algorithm we update π_g and λ_g with $\hat{\pi}_g = \frac{\sum_{i=1}^n \hat{z}_{ig}}{n}$ and $\hat{\lambda}_g = \frac{\sum_{i=1}^n \hat{z}_{ig} \hat{v}_{ig}}{\sum_{i=1}^n \hat{z}_{ig}}$, respectively. Then we use the Q-function to compute parameter updates for $\boldsymbol{\mu}_g$, $\boldsymbol{\Sigma}_g$, $\boldsymbol{\alpha}_g$ and η_g . The calculations for $\boldsymbol{\mu}_g$ yield the update

$$\hat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^n \hat{z}_{ig} \hat{v}_{ig} (\mathbf{x}_i - \mathbf{A}_g \boldsymbol{\eta}_{ig}) + \sum_{i=1}^n \hat{z}_{ig} \frac{(1 - \hat{v}_{ig})}{\eta_g} (\mathbf{x}_i - \sqrt{\eta_g} \mathbf{A}_g \tilde{\boldsymbol{\eta}}_{ig})}{\sum_{i=1}^n \hat{z}_{ig} \hat{v}_{ig} + \sum_{i=1}^n \hat{z}_{ig} \frac{(1 - \hat{v}_{ig})}{\eta_g}}, \quad (28)$$

and the update for $\boldsymbol{\Sigma}_g$ becomes

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_g &= \frac{1}{\sum_{i=1}^n \hat{z}_{ig}} \left[\sum_{i=1}^n \hat{z}_{ig} \hat{v}_{ig} \left\{ (\mathbf{x}_i - \boldsymbol{\mu}_g - \mathbf{A}_g \boldsymbol{\eta}_{ig})(\mathbf{x}_i - \boldsymbol{\mu}_g - \mathbf{A}_g \boldsymbol{\eta}_{ig})' \right. \right. \\ &\quad + \left. \mathbf{A}_g (\boldsymbol{\psi}_{ig} - \boldsymbol{\eta}_{ig} \boldsymbol{\eta}'_{ig}) \mathbf{A}'_g \right\} + \sum_{i=1}^n \hat{z}_{ig} (1 - \hat{v}_{ig}) \left\{ \frac{1}{\eta_g} (\mathbf{x}_i - \boldsymbol{\mu}_g - \sqrt{\eta_g} \mathbf{A}_g \tilde{\boldsymbol{\eta}}_{ig}) \right. \\ &\quad \times \left. (\mathbf{x}_i - \boldsymbol{\mu}_g - \sqrt{\eta_g} \mathbf{A}_g \tilde{\boldsymbol{\eta}}_{ig})' + \mathbf{A}_g (\tilde{\boldsymbol{\psi}}_{ig} - \tilde{\boldsymbol{\eta}}_{ig} \tilde{\boldsymbol{\eta}}'_{ig}) \mathbf{A}'_g \right\} \Big]. \end{aligned} \quad (29)$$

Recalling that \mathbf{A}_g is the diagonal matrix version of $\boldsymbol{\alpha}_g$, we obtain the update for skewness

$$\begin{aligned} \hat{\boldsymbol{\alpha}}_g &= \left[\sum_{i=1}^n \hat{z}_{ig} \hat{v}_{ig} (\boldsymbol{\Sigma}_g^{-1} \odot \boldsymbol{\psi}_{ig}) + \sum_{i=1}^n \hat{z}_{ig} \frac{(1 - \hat{v}_{ig})}{\sqrt{\eta_g}} (\boldsymbol{\Sigma}_g^{-1} \odot \tilde{\boldsymbol{\psi}}_{ig}) \right]^{-1} \\ &\times \left[\sum_{i=1}^n \hat{z}_{ig} \hat{v}_{ig} (\boldsymbol{\Sigma}_g^{-1} \odot \boldsymbol{\eta}_{ig} (\mathbf{x}_i - \boldsymbol{\mu}_g)') + \sum_{i=1}^n \hat{z}_{ig} \frac{(1 - \hat{v}_{ig})}{\eta_g} (\boldsymbol{\Sigma}_g^{-1} \odot \tilde{\boldsymbol{\eta}}_{ig} (\mathbf{x}_i - \boldsymbol{\mu}_g)') \right] \mathbf{1}_p. \end{aligned} \quad (30)$$

where the operator \odot denotes the Hadamard (elementwise) product of two matrices of the same dimension.

At the second CM-step of the ECM algorithm we update the contamination factor η_g . Looking at the Q-function in (25) we observe that there are two possible ways of updating η_g .

1. Use the R function `optim()` to find a maximum for Q , as η_g varies in the interval $(1, \eta_g^{\max})$, by considering

$$\begin{aligned} Q &= \sum_{i=1}^n \sum_{g=1}^G \hat{z}_{ig} (1 - \hat{v}_{ig}) \left\{ -\frac{p}{2} \log(\eta_g) - \frac{1}{2\eta_g} (\mathbf{x}_i - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g) \right. \\ &\quad + \left. \frac{1}{2\sqrt{\eta_g}} \left[(\mathbf{x}_i - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{A}_g \tilde{\boldsymbol{\eta}}_{ig}) + (\mathbf{A}_g \tilde{\boldsymbol{\eta}}_{ig})' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g) \right] \right\} \\ &\quad + \text{terms independent of } \eta_g. \end{aligned} \quad (31)$$

2. Solve the equation $\frac{\partial Q}{\partial \eta_g} = 0$ for η_g directly,

$$\begin{aligned}
0 &= \eta_g \left(-p \sum_{i=1}^n \hat{z}_{ig}(1 - \hat{v}_{ig}) \right) + \sum_{i=1}^n \hat{z}_{ig}(1 - \hat{v}_{ig})(\mathbf{x}_i - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g) \\
&\quad - \eta_g^{1/2} \sum_{i=1}^n \hat{z}_{ig}(1 - \hat{v}_{ig}) \left[(\mathbf{x}_i - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{A}_g \tilde{\boldsymbol{\eta}}_{ig}) + (\mathbf{A}_g \tilde{\boldsymbol{\eta}}_{ig})' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g) \right],
\end{aligned} \tag{32}$$

which is a quadratic equation of the form $S\eta_g + T\eta_g^{1/2} + U = 0$ that can be easily solved for η_g (where S, T, U are constants).